Show Me How To Revise: Improving Lexically Constrained Sentence Generation with XLNet Appendix

Language Models

For LSTM-based language models, both the forward and backward language models have two LSTM layers. The embedding size and hidden size are 256. In addition, we selected 50,000 most frequent tokens of the training set as the vocabulary for LSTM-based models. During the training process, we set the dropout to 0.2 and the learning rate to $1e-4$.

For XLNet-based language models, we used the pre-trained XLNet (base-cased version) model, which has 12 self-attention layers with 110M parameters. The vocabulary size for XLNet is 32,000. We fine-tuned XLNet on the training set with learning rate $lr = 1e-5$. We trained all language models on the training set until no improvement on the validation set. We selected the best checkpoint with the lowest validation loss. NLL results of the trained language models on the One-Billion-Word validation set are shown in Table 5.

Models	Forward	Backward
LSTM-based	4.272	4.448
XLNet-based	3.076	3.071
GPT-2 (small)	3.980	-

Table 5: NLL of different language models on the One-Billion-Word validation set. GPT-2 is used to evaluate the quality of the generated sentences. GPT-2 is the pre-trained model without any fine-tuning.

Classifier

We created the synthetic dataset for the copy, replacement, insertion, and deletion actions. Similar to the replacement action, we also resorted to two approaches to create synthetic data for the deletion action. To create synthetic data for the deletion action, we need to insert some tokens in selected sentences, where the inserted tokens are randomly sampled from the vocabulary or predicted by the masked language model, i.e., XLNet. Both the training and validation sets are created with masked LM and random methods. We created 36M sentences as the training set (30M sentences are created with the masked LM method, and 6M sentences are created with the random method) and 1.8M sentences as the validation set (1.5M sentences are created with the masked LM method, and 0.3M sentences are created with the random method). We fine-tuned the pre-trained XLNet (base-cased version) model on the synthetic training set with learning rate $lr = 1e-5$ for two epochs. We used precision, recall, and F1 score to evaluate the fine-tuned classifier on the validation set. During training, we chose the checkpoint with the best performance on the validation set on the macro-average F1 score. From Table 7, we can see that the fine-tuned classifier achieves high performance on the random validation set and performs slightly worse on the masked LM validation set, mainly because it is more challenging to infer labels for the masked LM validation set.

In our experiments, we only leveraged the fine-tuned classifier to guide MCMC sampling to insert or replace tokens without using the deletion action since the deletion action makes the model hesitate to move forward. For example, the model may insert a token and then delete it later, because at the beginning, both insertion and deletion actions have high probabilities. We empirically found that the incomplete sentence can be refined by iteratively inserting tokens. Therefore, only using the replacement and insertion actions can generate plausible sentences.

To better understand the classifier's function, we showed the learned prior of four candidate sentences in Figure 5. In the subfigure (a), we can see the classifier tells us we should insert some tokens before 'film' and '<EOS>'. The other tokens should maintain unchanged. This prediction is consistent with our intuition.

N-gram Repetition

We measured whether a sentence contains n-gram repetitions based on three rules. Firstly, if a unigram appears more than three times in a sentence, we regarded the sentence contains a repetition. Similarly, if a bigram appears more two times or a trigram appears more than one time in a sentence, the sentence is also considered as containing a repetition.

Metrics	Repetition (↓)			
Models	k=1	k=2	k=3	k=4
Human Reference	3.4%	4.3%	4.3%	4.3%
sep-B/F	32.4%	-	-	-
asyn-B/F	39.6%	-	-	-
GBS	21.2%	24.8%	21.6%	26.1%
CGMH	0.5%	0.2%	0.2%	0.2%
L-MCMC	0.1%	0	0	0.2%
L-MCMC-C	0.3%	0.7%	0.4%	0.2%
X-MCMC	0	0	0	0
X-MCMC-C	0.1%	0.2%	0.2%	0

Table 6: The percentage of sentences containing n-gram repetitions. (k denotes the number of lexical constraints.)

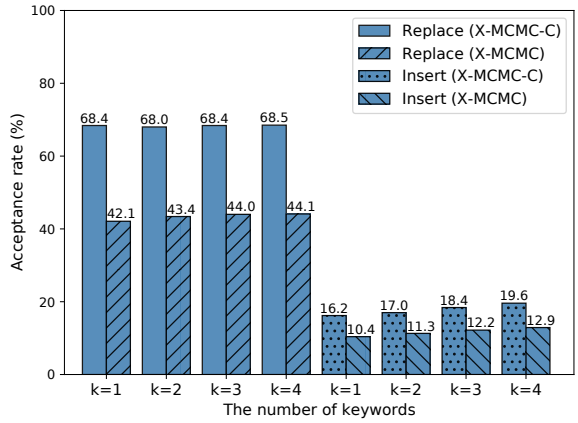


Figure 4: Acceptance rates vs. the number of constraints k.

We showed the percentage of sentences containing n-gram repetitions in Table 6. All MCMC-based models (CGMH, L-MCMC, L-MCMC-C, X-MCMC, and X-MCMC-C) are run for 200 steps. We can see that sentences generated by beam search-based models (sep-B/F, asyn-B/F and GBS) tend to get stuck in repetitions.

Acceptance Rates of XLNet-based Models

We show the acceptance rates of XLNet-based models (X-MCMC and X-MCMC-C) in Figure 4. Compared with X-MCMC, X-MCMC-C has much higher acceptance rates, consistent with their LSTM-based counterparts.

Sentences Generated with Lexical Constraints

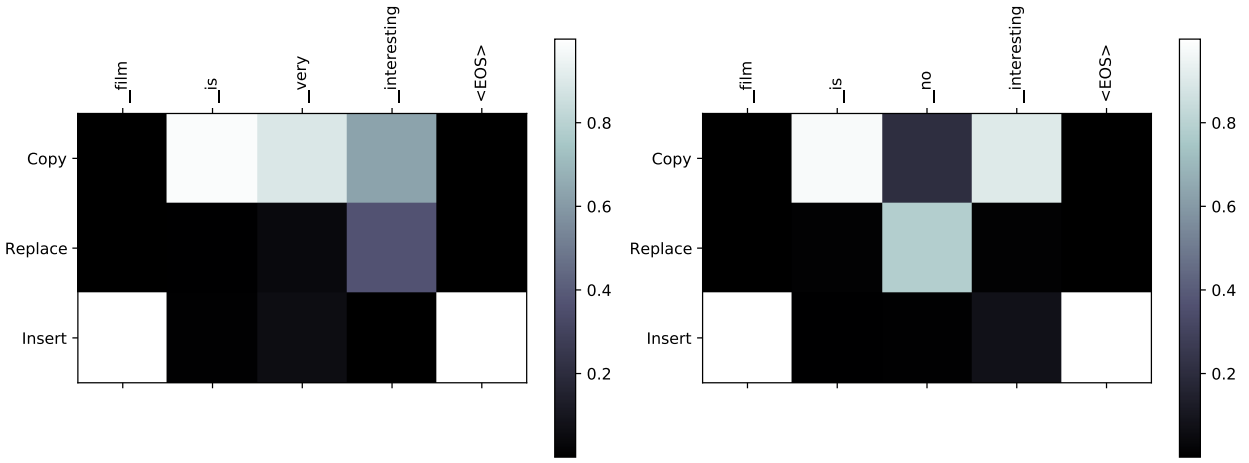
We show some sentences generated by different models with lexical constraints in Table 8. All MCMC-based models (CGMH, L-MCMC-C, and X-MCMC-C) are run for 200 steps.

Text Infilling

We show some sentences generated by different models with different templates in Table 9.

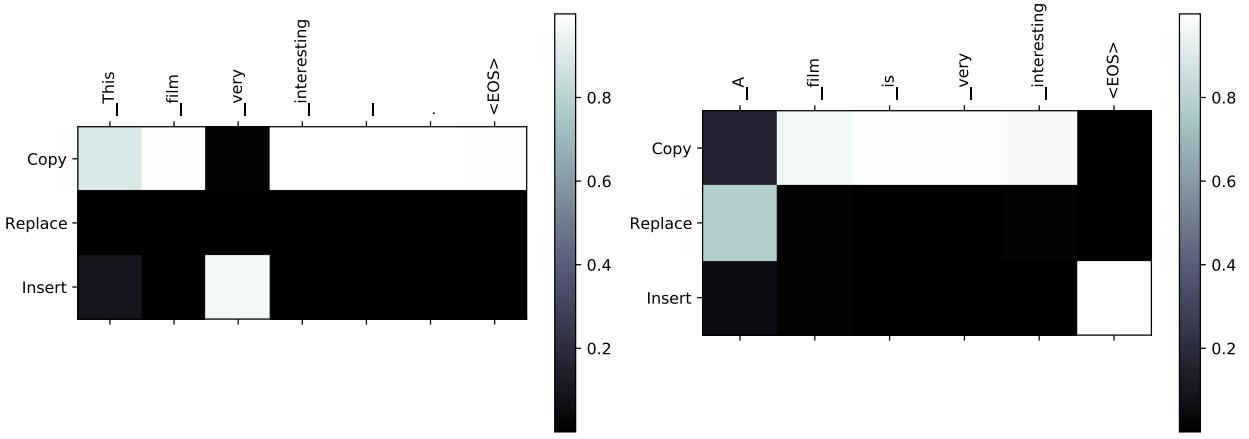
	Masked LM validation set			Random validation set			Whole validation set		
Labels	P (↑)	R (↑)	F1 (↑)	P (↑)	R (↑)	F1 (↑)	P (↑)	R (↑)	F1 (↑)
Copy	0.978	0.993	0.986	0.989	0.996	0.993	0.980	0.994	0.987
Replacement	0.772	0.462	0.578	0.872	0.801	0.835	0.795	0.519	0.628
Insertion	0.951	0.893	0.921	0.947	0.894	0.920	0.951	0.893	0.921
Deletion	0.809	0.684	0.741	0.875	0.860	0.868	0.821	0.714	0.764
Macro-average	0.878	0.758	0.807	0.921	0.888	0.904	0.887	0.780	0.825

Table 7: Results of the classifier on the synthetic validation sets. "P" and "R" denote precision and recall. The random validation set is created by replacing some tokens with random tokens. The masked LM validation set is created by the masked LM method. The whole validation set combines the random and masked LM validation sets.



(a) Example 1

(b) Example 2



(c) Example 3

(d) Example 4

Figure 5: The learned prior of four different candidate sentences given by the classifier. The x-axis and y-axis of each plot correspond to the tokens in the candidate sentence and the actions.

Constraints	**view**
Human	Victims commissioner Brendan McAllister has asked the Policing Board for its **view** on the transfer .
sep-B/F	He said : " There is a point of **view** , " he said in a statement .
asyn-B/F	He said : " There is a point of **view** , " he told BBC Radio 4 's Today programme .
GBS	" It 's a good thing to do , " he said , adding that he did not have a **view** of the situation .
CGMH	If this case ends up or next to the **view**
L-MCMC-C	What is the world 's **view** of the global economy ?
X-MCMC-C	The consensus **view** is that this has not been achieved .
Constraints	**promoting**, **energy**
Human	They ranged from **promoting** clean **energy** and overseas infrastructure projects to attracting medical tourists and foreign firms .
GBS	" We are **promoting energy** efficiency , " said <UNK> <UNK> , a professor of economics at the University of California , Berkeley .
CGMH	But it is now widely criticized for **promoting** the **energy**
L-MCMC-C	He is also committed to **promoting** innovation in renewable **energy** projects .
X-MCMC-C	They are expected to talk about curbing carbon emissions and **promoting energy** efficiency .
Constraints	**experience**, **feels**, **dream**
Human	" This whole **experience feels** like a **dream** come true . ...
GBS	" It **feels** like it 's a **dream** , " he said , adding that he had no **experience** in the world 's history .
CGMH	In his **experience** , he **feels** he can 't find his own **dream**
L-MCMC-C	The whole **experience feels** like a very different story – a man 's **dream** .
X-MCMC-C	The whole **experience feels** like a **dream** to me .
Constraints	**seized**, **movement**, **party**, **taken**
Human	They have **seized** on a common claim : A dangerous fringe **movement** , the tea **party** , has **taken** over the Republican Party .
GBS	The **party** has **taken** part in a series of meetings with the opposition **movement** , which has been **seized** in recent years .
CGMH	He has also **seized** his head off his **movement** toward his own **party** and has **taken**
L-MCMC-C	The MDC has **seized** power , and the Islamist **movement** 's main opposition **party** has been **taken** over .
X-MCMC-C	The two **seized** rebel **movement** leaders and the local **party** leader were also **taken** to police stations .
Constraints	**talking**, **something**, **analysts**, **planning**
Human	But it is unlikely he 'll be **talking** to victims , **something** Vatican **analysts** had suggested was a possibility in the early **planning** stages of the U.S. trip .
GBS	" I think it 's **something** we 're **talking** about , " said <UNK> , who is **planning** to meet **analysts** on Wednesday .
CGMH	After **talking** for **something** that happened , some **analysts** think it is **planning**
L-MCMC-C	The first meeting came on Wednesday , after **talking** about **something** that some **analysts** were **planning** .
X-MCMC-C	They are **talking** about **something** that some **analysts** have been **planning** for for several years .

Table 8: Sentences generated by different models with lexical constraints on the One-Billion-Word test sets. ("Human" refers to the human reference.)

Template	Moving ____ from ____ induce problems ____ depression ____
Human Reference	Moving **them** from **person to person , home to home , can** induce problems **such as anxiety and** depression **.**
FLM	Moving **forward** from **the end of the year , he said that** induce problems **could be used for** depression **.**
BLM	Moving **away** from **embryonic stem cells , which can be used to** induce problems **, heart disease and** depression **.**
Bayesian MCMC	Moving **down** from **the top that you say is that they can** induce problems **as a result of** depression **.**
TIGS (strategy 1)	Moving **away** from **actor David James Jones to work with music to** induce problems **in modern film classic** depression **shows**
TIGS (strategy 2)	Moving **away** from **the United States , the United States has been** induce problems **in the past major** depression **.**
L-MCMC	Moving **away** from **the health care system that does not** induce problems **can cause a serious** depression **.**
L-MCMC-C	Moving **away** from **the U.S. banking system can also be used to** induce problems **related to global economic** depression **.**
Template	Opponents ____ say U.S. ____ suffer under the climate bill ____ trade ____ changes ____
Human Reference	Opponents **of the tariff** say U.S. **manufacturing would** suffer under the climate bill **regardless of** trade **policy** changes **.**
FLM	Opponents **of the bill** say U.S. **banks are** suffer under the climate bill **.** <**EOS**> trade **and** changes **to**
BLM	Opponents **are expected to** say U.S. **does not** suffer under the climate bill **, to** trade **the** changes **.**
Bayesian MCMC	Opponents **are expected to** say U.S. **lawmakers might** suffer under the climate bill **in the** trade **of** changes **.**
TIGS (strategy 1)	Opponents **Michael Jackson supporters** say U.S. **car owners** suffer under the climate bill **plan to** trade **the** changes **government**
TIGS (strategy 2)	Opponents **of the bill** say U.S. **banks will** suffer under the climate bill **, which** trade **on** changes **in**
L-MCMC	Opponents say U.S. **citizens will** suffer under the climate bill **and also** trade **the** changes **.**
L-MCMC-C	Opponents say U.S. **companies will** suffer under the climate bill **and make** trade changes **.**
Template	The doctor said ____ the newspaper reported .
Human Reference	The doctor said **it was unclear whether she will walk again , even though some people can recover from such injuries ,** the newspaper reported .
FLM	The doctor said **there was no evidence that he had been involved in the incident .** <**EOS**> **.** <**EOS**> **.** <**EOS**> **,** the newspaper reported .
BLM	The doctor said **the company posted a net profit of more than $ 1 billion by the end of last year ,** the newspaper reported .
Bayesian MCMC	The doctor said **that a 21-year-old man whose body was not taken to the hospital would have suffered from stab wounds ,** the newspaper reported .
TIGS (strategy 1)	The doctor said **two clothing company Medical Center 's National Park Service found pictures of pop star Michael Jackson Ray Allen 18** the newspaper reported .
TIGS (strategy 2)	The doctor said **the boy was not a man who was not involved in the incident .** <**EOS**> **, with the statement** the newspaper reported .
L-MCMC	The doctor said **he was not too worried ,** the newspaper reported .
L-MCMC-C	The doctor said **it was a possible suicide attack , but it was not immediately clear what happened at the time ,** the newspaper reported .

Table 9: Example outputs of different models with different templates on the One-Billion-Word test sets.